

Genome-level analysis of genetic regulation of liver gene expression networks

Daniel M. Gatti

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree of Masters of Science in the
Department of Environmental Sciences and Engineering.

Chapel Hill

2007

Approved by:

Ivan Rusyn, M.D. Ph.D.

David Threadgill, Ph.D.

Fred Wright, Ph.D.

Abstract

Daniel M. Gatti : Genome-level analysis of genetic regulation of liver gene
expression networks

(Under the direction of Ivan Rusyn, M.D., Ph.D.)

The liver is the primary site for metabolism of nutrients, drugs and chemical agents. While metabolic pathways are complex and tightly regulated, genetic variation among individuals, reflected in variation in gene expression levels, introduces complexity into research on liver disease. This study aimed to dissect genetic networks that control liver gene expression by combining large-scale quantitative mRNA expression analysis with genetic mapping in a reference population of BXD recombinant inbred mouse strains for which extensive SNP, haplotype and phenotypic data is publicly available. We profiled gene expression in livers of naive mice of both sexes from C57BL/6J, DBA/2J, B6D2F1, and 37 BXD strains using Agilent oligonucleotide microarrays. This data was used to map quantitative trait loci (QTLs) responsible for variation in expression of about 19,000 transcripts. We identified polymorphic cis- and trans-acting loci, including several loci that control expression of large numbers of genes in liver.

Acknowledgements

This work would not have been possible without the collaboration the following individuals and institutions. I would like to thank :

Dr. Akira Maki¹ for isolating cDNA and running the microarray experiments,

Dr. Lu Lu^{3,6} for housing, feeding and sacrificing the mice in the study,

Dr. Robert Williams³, Dr. Ken Manly³ & Yanhua Qu³ for WebQTL and project advice,

Dr. Elissa Chesler², Dr. Michael Langston⁴, Dr. Roumyana Kirova² and Andy Perkins⁴ for batch normalization of the microarray data, QTL analysis and the para-clique results,

Dr. David Threadgill⁵ for invaluable direction in the course of analyzing the data,

Oksana Kosyk¹ for running all RTPCR experiments and

Greg Hannon and Elizabeth Murchison from Cold Spring Harbor Labs for providing liver samples from *Dicer1* heterozygous animals.

¹Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599,

²Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831,

³Department of Anatomy and Neurobiology, University of Tennessee Health Science Center, Memphis, TN 38163,

⁴Department of Computer Science, University of Tennessee, Knoxville, TN 37996, and

⁵Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

⁶Key Laboratory of Nerve Regeneration, Nantong University, Nantong 226001, PR China

Table of Contents

	Page
List of Tables.....	v
List of Figures.....	vi
List of Abbreviations.....	vii
Chapter 1 : Literature Review.....	8
Chapter 2 : Introduction.....	12
Chapter 3 : Experimental Methods.....	14
Chapter 4 : Results and Discussion.....	18
Chapter 5 : Conclusions and Future Directions.....	35
References.....	43

List of Tables

Table	Page
1 : BXD Microarray Experiment Batch Design.....	39
2 : BXD strain information	41

List of Figures

Figure	Page
1 : WebQTL interval mapping reveals genetic control of gene expression.....	19
2 : Genome-wide clustering of the genetic control of gene expression in liver.....	21
3 : Tissue-specific transcriptome maps reveal differences and similarities in genetic regulation of gene expression.....	25
4 : Chr12 locus is a master-regulator of gene expression in mouse liver.....	27
5 : <i>In silico</i> discovery of gene expression to phenotype correlations using WebQTL.....	32
6 : WebQTL-assisted strain selection for phenotypic profiling.....	34

List of Abbreviations

AIL - Advanced intercross line

BXD - Recombinant inbred cross between C57BL/6J and DBA/2J

FDR - False discovery rate

GO - Gene Ontology

LRS - Likelihood ratio statistic

QTL - Quantitative trait loci

RI - Recombinant inbred

CHAPTER 1

Literature Review

With the recent sequencing of entire mammalian genomes^{1,2}, the possibility exists that the genetic basis of liver injury can be unraveled. Several high density genotyping projects are underway to sequence panels of inbred strains at densities of ~150,000 and ~8.3 loci^{3,4}, providing an unprecedented density of genomic information. This detailed sequence data can be combined with high throughput gene expression microarrays to examine the effect of genetics on constitutive levels of gene expression⁵⁻⁷. The combination of such high dimension data sets can offer insight into complex mechanisms of liver disease and toxicity⁸.

Liver injury is the primary reason that prospective drugs are pulled from clinical testing and such idiosyncratic liver responses account for up to 50% of liver failure cases⁹. While the drug in testing may behave as designed in most of the population, some proportion experiences serious and unpredicted toxicity. This toxicity is typically not observed in animal models and has a low incidence in the general population¹⁰. These properties make idiosyncratic liver injury extremely difficult to predict. Further, if such reactions occur when patients are intentionally exposed to drugs, it is possible that environmental xenobiotics can have the same effects. While most of the population does not experience an adverse response to common environmental exposures, a small percentage may experience an idiosyncratic response. An increased understanding of

the mechanisms of toxicity in the liver will allow us to better predict and prevent such incidents.

All human populations are exposed to environmental toxins and toxicants. However, the response to such exposures is not uniform. While the majority of the population may experience exposure and suffer no measurable injury, a minority may experience severe toxicity. The reasons for such differential responses are many. Exposure dose and duration, lifestyle and nutritional choices, genetic differences and other variables can all affect the response. In this study, the focus is specifically on the genetic differences that influence the differential response to toxic insult.

In order to study these genetic differences, mouse models are used. Mice have been used in the study of many chemicals including 1,1,2-trichloroethylene¹¹, cadmium¹² and 1,3-butadiene¹³. Mice are preferable to simpler organisms because they are easy to house and feed, breed well and are mammalian systems with more similarity to humans than fish or insects. They are also preferable to *in vitro* systems because cells in culture may or may not respond identically to cells in the body. Further, *in vivo* systems retain the complex signaling interactions between disparate cell types within and between organs^{14,15}.

In order to study the effect of genetic background on gene expression, mice with a controlled genotype are used. Inbred strains are homozygous at all genomic loci and whose individuals have identical genotypes. These strains are developed through extended sibling mating for over 20 generations. This process leads to increased homozygosity over at least 99% of the genome. The use of inbred strains in genetics is well established and some strains have been bred and used in laboratory experiments for almost 100 years¹⁶. The decreased genetic diversity within a single inbred strain leads to greatly reduced phenotypic variance, reducing the number of mice needed to detect statistically significant phenotypic differences. The stable genotype is also invaluable for

reproducing phenotypic measurements in different laboratories. Environmental factors can be varied in a controlled manner without additional genotypic variation¹⁷. It is important to note that while the phenotypic variation within an inbred strain is reduced, the phenotypic variation among a panel of inbred strains can remain vast¹⁸.

The progeny of inbred mice can be bred in such a way that the genomes are permuted in a controlled manner. Such recombinant inbred (RI) mice are inbred mice whose progenitors are themselves inbred mice. These mice are produced by mating two inbred parents (i.e. C57BL/6J & DBA/2J) to produce an F1 generation. The F1 litter is sibling mated to produce an F2 generation. Each F2 generation is then sibling mated for over 20 generations to produce a unique RI line. The naming of these lines uses an 'X' between a letter from each of the two progenitor strains. For example, the C57BL/6J and DBA/2J derived RI line is called 'BXD.' The advantage of using RI strains is that they provide a known ancestry and a controlled mixing of genotypes while retaining the homozygosity and reduced intrastrain phenotypic variance of inbreds.

Quantitative Trait Locus (QTL) mapping has been used to associate a specific genotype with the variation in a single measured phenotype like high density lipoproteins¹⁹ and ethanol tolerance²⁰. At each locus in a segregating population, a model is fit which estimates the likelihood that this locus explains the variation in phenotype versus the likelihood that there is no genotypic effect on the phenotype.

Interval mapping²¹ is a variation on QTL mapping which uses maximum likelihood estimation. At each marker, the likelihood that the marker is associated with the phenotype over the likelihood that the marker is associated with no genotype is calculated. This is the Likelihood Ratio Statistic (LRS). Between markers, the exact genotype is unknown. Because of this missing data, the expectation maximization algorithm is used to estimate the genotype based on recombination frequencies. The

result is a QTL plot of the entire genome with LRS scores indicating the strength of association between the phenotype and each genomic location.

While QTL mapping was initially used with a single phenotype, the arrival of microarray technology allows for the measurement of thousands of phenotypes simultaneously. Beginning with a study in yeast²², QTL mapping has been done with gene expression as the phenotype. In such a study, the genomic loci responsible for variation in gene expression can be used to infer regulatory control. While such a study is not conclusive, it can be used to narrow the potential regulatory candidates, generate hypotheses for further testing and construct regulatory networks *in silico*.

This approach has been applied in the mouse liver⁵, brain⁶ and hematopoietic stem cells⁷. The liver study⁵ used an F₂ population derived from C57BL/6J and DBA/2J to study the relationship of liver gene expression to fat pad mass. The others^{6,7} looked at constitutive gene expression in a panel of BXD RI mice to infer the regulation of basal gene expression. These studies were significant in increasing our understanding gene regulatory networks including the existence of master regulator loci, clusters of co-regulated genes and the association of gene expression with previously measured behavioral traits.

Missing from the current literature is a study of constitutive gene expression in the mouse liver using a panel of RI strains. Such a study would lay a foundation of basic research for liver toxicologists and provide a resource which can be used to associate genes expression with phenotype measurements. It could be used to find clusters of co-regulated genes that may not have been otherwise associated in previous work and to build regulatory networks for genes relevant to xenobiotic metabolism. Lastly, such a data set could be used for intelligent strain selection when a knockout model is otherwise unavailable.

CHAPTER 2

Introduction

The maturation of gene expression technology has opened the door to the exploration of the genetics of gene expression^{5,6,22}. Microarrays allow for the concurrent measurement of thousands of transcripts with the resultant genomic data being increasingly used to improve biological interpretation of data from mechanistic research. Phenotypic anchoring of observed phenotypes to gene expression changes has proven useful in uncovering the molecular mechanisms that lead to liver injury^{23,24}. Such experiments connect the variation in transcript expression to phenotypes. However, they do not lead to detailed gene expression networks where the expression of one gene is found control the expression of another.

Recombinant Inbred (RI) mice are created by crossing two parental strains followed by sib-mating for over 20 generations²⁵. Strains created this way have the advantage of being homozygous at almost every location along the genome. As such, each representative of an RI line will have limited phenotypic variation within that line, but the variation between lines is usually vast. RI panels are widely used to determine genotype-phenotype associations using QTL mapping techniques²¹. The relationships between phenotypes and genotypes are calculated using a likelihood ratio statistic (LRS), which is a measure of the probability that a given genetic marker explains the variation in the phenotype. When mRNA levels are used as the phenotype, regions of the genome with a high LRS are likely to contain genes that control the expression of the gene

transcript being profiled; this process is referred to as expression Quantitative Trait Loci (eQTL) mapping ^{5,26}.

The BXD panel of RI strains was created from the C57BL/6J and DBA/2J parental strains ^{27,28}. These two parental inbred strains are known to exhibit widely different phenotypes in response to a number of exposures. Thus, BXD mice have been a useful tool for elucidation of the genetic control of certain diseases. For example, the BXD lines have been used to study alcohol preference and tolerance ²⁹, alcohol metabolism ³⁰, responsiveness to aromatic hydrocarbons ³¹, N,N-diethylnitrosamine induced hepatocarcinogenesis ³² and diabetes and atherosclerosis ^{33,34}. Recently, BXD mice were used for eQTL studies that elucidated the genetics of gene expression in the brain ⁶ and hematopoietic stem cells ⁷.

The genetic basis and networks that control gene expression in the liver are not well delineated and improving our understanding of these pathways and controlling loci will advance knowledge of physiological and pathophysiological changes in this organ. Here, eQTL mapping was applied to data from genome-wide microarray profiling of liver gene expression in the naïve state of the parental C57BL/6J and DBA/2J strains, the F1 cross and 37 BXD strains. Using this approach, potential key regulators of gene expression in the liver were identified. The comparison of the liver and brain transcriptome maps demonstrated tissue-specific differences in regulation of gene expression. Finally, we demonstrate how the data collected in this study may be used to infer genotype-phenotype correlations, generate testable hypothesis, and identify strains that may differ in responsiveness to xenobiotics based on the genetically-determined differences in expression of the key genes.

CHAPTER 3

Experimental Methods

Animals and Tissues. BXD1 through BXD42 mice are original RI strains available from the Jackson Laboratory (Bar Harbor, ME). BXD43 through BXD100 lines were generated using ninth or tenth generation Advanced Intercross Line (AIL) progenitors. AILs are generated by breeding two inbred parents (here is C57BL/6J and DBA/2J) and crossing their offspring so as to minimize inbreeding and maximize recombination events at each generation ²⁸. Mice were maintained at 20-24°C on a 14/10 hr light/dark cycle in a pathogen-free colony at the University of Tennessee-Memphis. Animals were fed a 5% fat Agway Prolab 3000 rat and mouse chow and given tap water in glass bottles. Strain details are provided in Supplemental Table 1. Mice were raised to between 54 and 177 days (mean 70 days) of age. Liver tissues were collected following sacrifice by cervical dislocation. The whole liver was removed immediately, and placed in five volumes of RNAlater (Ambion, Austin, TX) at 4°C overnight, before removing from RNAlater and storing at -80°C until processing. All animal studies for this project were approved by Animal Care and Use Committee at the University of Tennessee-Memphis.

RNA Isolation. Total RNA was isolated from liver samples (~30 mg) using the RNeasy mini kit (Qiagen, Valencia, CA) as detailed by the manufacturer. RNA quality and quantity was determined spectrophotometrically from the absorbance at 260 nm and 280 nm. Aliquots of RNA samples were frozen at -80°C until microarray analysis. For each microarray, RNA was pooled from 2-3 mice of the same sex and strain.

Microarray Analysis. One μg of total RNA from 3 individual mouse liver samples of the same sex and strain was pooled, amplified and labeled with a fluorescent dye (Cy5) and a common reference - pooled mouse tissues (equal amounts of RNA from liver, kidney, lung, brain and spleen) mRNA isolated from the livers of 100 male C57BL6/J mice [see ³⁵ for details] - were amplified and labeled with Cy3 using Agilent Technologies (Palo Alto, CA) Low RNA Input Linear Amplification labeling kit following the manufacturer's protocol. The quantity of the resulting fluorescently labeled cRNA was measured using a Nanodrop ND-100 spectrophotometer (Nanodrop Technologies, Wilmington, DE) and its integrity assessed using an Agilent Bioanalyzer. Equal amounts of Cy3 and Cy5-labeled cRNA (750 ng) from the individual animals and from the pooled control, respectively, were hybridized to an Agilent Mouse Oligo Microarray (~21,000 features, catalogue# G4121) for 17 hours at 65°C. The hybridized microarrays were then washed and scanned using an Agilent G2565BA scanner. Data were extracted from the scanned image using Agilent Feature Extraction software version 6.1. A total of 122 arrays were run in 8 batches. The samples were semi-randomly distributed throughout the batches prior to microarray analysis in order to separate sexes and strains, and to minimize between- and within-batch bias. Technical and biological replicates were run both within each batch and between batches. The microarray data was deposited in the UNC Microarray Database and extracted using Log2 ratios of the mean red channel intensity over the mean green channel intensity. This was followed by LOWESS normalization to remove the intensity dependent dye bias ³⁶. Neither the genes nor the arrays were centered. Inter-batch normalization was carried out using a nested ANOVA mixed model with samples within each batch crossed with sex and strain.

QTL Analysis and WebQTL. QTL linkage mapping was carried out using the QTL Reaper software package (qtlreaper.sourceforge.net). One thousand permutations of the strain labels were performed to estimate the genome wide p-value ³⁷. Liver expression data was deposited in WebQTL (www.genenetwork.org) which is a web-based resource for exploring gene expression and phenotype interactions. WebQTL was used to produce interval maps for specific genes.

Transcriptome Map and Whole Genome QTL Clustering. The transcriptome map of the liver was produced using the R statistical package using the output data from QTL Reaper which consisted of the maximum Likelihood Ratio Statistic (LRS) value for each transcript on the microarray as well as a permutation derived p-value (number of permutations = 1000). The p-value threshold was applied at a 25% False Discovery Rate ³⁸. The QTL data for all informative markers and 18,716 transcripts on the G4121A microarray was used to perform hierarchical clustering of the QTL data. Transcripts were clustered using each 2,325 marker vector of LRS values. The distance metric used for complete (maximum distance) hierarchical clustering was "1 – Pearson correlation" and colored by normalized LRS value.

Transcription Factor Analysis. Three web-based tools were used to search for possible transcription factor binding sites in candidate loci: the National Cancer Institute's Advanced Biomedical Computing Center promoter analysis tool (grid.abcc.ncifcrf.gov/promoters.php), oPossum version 1.3 ³⁹ and PAINT ⁴⁰. In the first two cases, transcripts were divided based on increased expression correlating with the C57BL/6J of DBA/2J allele. The transcripts were submitted in 4 groups: 1) high expression with the C57BL/6J allele and an LRS of 30 or greater, 2) high expression with the C57BL/6J allele and an LRS of 40 or greater, 3) high expression with the DBA/2J

allele and an LRS of 30 or greater and 4) high expression with the DBA/2J allele and an LRS of 40 or greater. With PAINT, all transcripts were submitted as one list, but a gene cluster file was also submitted that clustered the C57BL/6J high and DBA/2J high transcripts into separate clusters.

Quantitative Real Time PCR analysis. In order to test the hypothesis that *Dicer1* might be the regulatory gene at the distal chromosome12 locus which controls the 111 transcripts, female *Dicer1* wild type and heterozygote mice ⁴¹ were obtained. No knockout is available because *Dicer1* is embryonic lethal. 10 genes that are trans-regulated at this locus were selected (*Abhd1*, *Dhrsx*, *Hnf4g*, *Met*, *Neurog3*, *Olfr656*, *Pms1*, *Runx3*, *Rqcd1*, *Sel1h*). 9 genes were selected that are known to be expressed in the murine liver but do not have a significant or suggestive QTL on distal chromosome 12 (*Abcb11*, *Abcc2*, *Abcd3*, *Cyp2e1*, *Cyp3a11*, *Cyp7a1*, *Cyp8b1*, *Ppar1*, *Rara*). RNA isolation was carried out using the Qiagen (Valencia, CA) RNeasy Mini Kit 50. RNA concentration and quality was checked using the Nanodrop (Wilmington, DE) ND-1000 spectrophotometer at the 260 and 280 nm wavelengths. 20 µg of RNA were used to produce cDNA using the Applied Biosystems Inc. (Foster City, CA) High Capacity cDNA Archive Kit. The Stratagene (La Jolla, CA) FullVelocity QYBR Green QPC Master Mix was used to perform the RTPCR and the plates were run on a Stratagene Mx3000P instrument.

CHAPTER 4

Results and Discussion

Genetic control of gene expression in liver

Using web-based eQTL mapping tools and data collected in this study, regulatory loci controlling each liver transcript can be easily visualized in WebQTL (www.genenetwork.org). Figure 1 shows examples of the three types of expression control found in the liver. A transcript with a maximum locus of control near (+/- 20 Mb) the genomic location of the transcript itself is considered to be *cis*-regulated⁴². This implies some mechanism of control near the gene itself; perhaps a polymorphism in the promoter region. An example of a strong *cis*-regulated transcript is the cytochrome P450, family 3, subfamily a, polypeptide 13 (*Cyp3a13*) gene, located on distal chromosome 5 (Figure 1a). A transcript with a maximum locus of control far from the gene location is *trans*-regulated. This implies regulation by another gene like a transcription factor. ADP-ribosyltransferase 5 (*Art5*) gene, located in the middle of chromosome 7, is an example of a *trans*-regulated gene whose expression levels are determined by polymorphisms on chromosome 3 (Figure 1b). Interestingly, many liver-expressed transcripts are regulated by multiple loci. The interleukin 21 receptor (*Il21r*) gene, located on distal chromosome 7, is a good example (Figure 1c). *Il21r* has two significant QTL peaks - one on chromosome 7 proximal to the gene itself and one on chromosome 14.

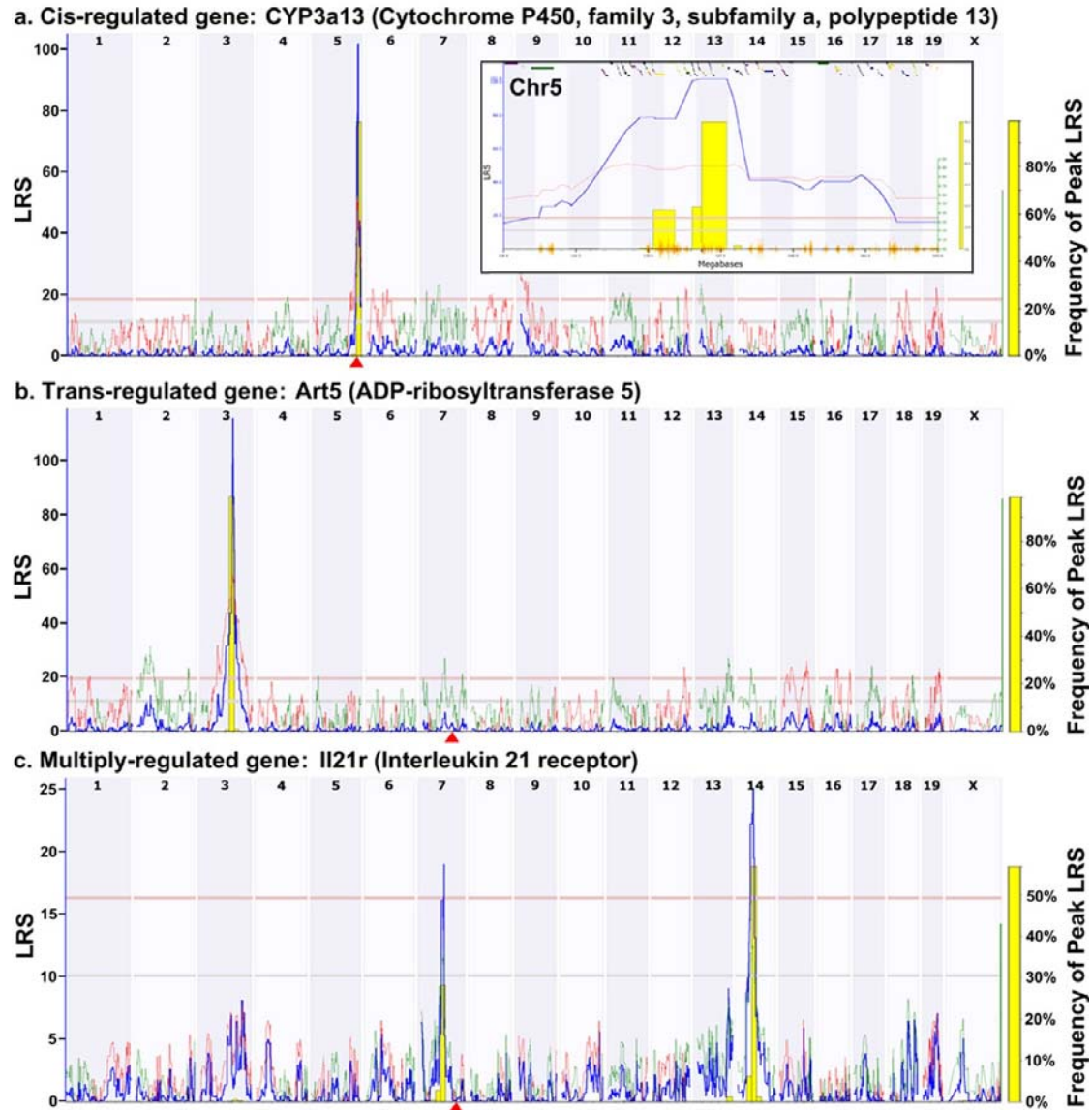


Figure 1. WebQTL interval mapping reveals genetic control of gene expression. a, An example of a cis-regulated gene (*Cyp3a13*) where the QTL is co-located with the gene. The horizontal axis displays the mouse genome. The vertical axis displays the Likelihood Ratio Statistic (LRS). The gene location is shown by the red triangle. The red horizontal line indicates a significant level of association as determined by permutation analysis. The grey horizontal line below it represents a suggestive association. The blue line displays the LRS along the genome and the yellow bars are the results of a bootstrap analysis. The inset shows a zoomed-in view of the *Cyp3a13* QTL on Chr5 (130-145Mb) and illustrates the features of WebQTL. The red line indicates that C57BL/6J alleles at the marker increase expression of the gene. The colored boxes along the top represent known genes which, when selected, will take the user to the NCBI EntrezGene entry for that gene. A SNP track is displayed along the bottom as an orange seismogram. **b,** An example a trans-regulated gene (*Art5*). **c,** An example of a gene (*Il21r*) that is regulated by multiple loci.

To visualize the patterns of genetic control of gene expression on a genome-wide level, the 18716 annotated transcripts were clustered using the LRS vector for each transcript (Figure 2a). As expected, the majority of the transcripts in liver are independently regulated; however, several distinct patterns emerge. Specifically, there are a number of clusters of transcripts that all share a common maximal QTL as well as clusters that are co-regulated by a complex set of common loci. We refer to clusters regulated by only one strong QTL as “simple” QTL clusters and ones regulated by multiple loci as “complex” clusters.

Chromosome 8 contains a simple cluster of transcripts that all have a strong maximum QTL (mean LRS = 47.5, Figure 2b). Of the 27 transcripts in this cluster, 26 are located on chromosome 8 at the same location as the maximum QTL which indicates that this cluster contains predominantly *cis*-regulated genes or perhaps is due to a strain-specific difference in a regional transcriptional enhancer. Interestingly, the presence of one of the parental (C57BL/6J or DBA/2J) allele at this locus strongly affects expression of these genes (Figure 2b, yellow-red correlation plot).

Chromosome 12 also contains a simple cluster that consists of a set of transcripts with a maximal QTL (mean maximum LRS = 42.6) at the distal region of the chromosome (Figure 2c). However, only 5 of the 111 transcripts in this cluster are *cis*-regulated and

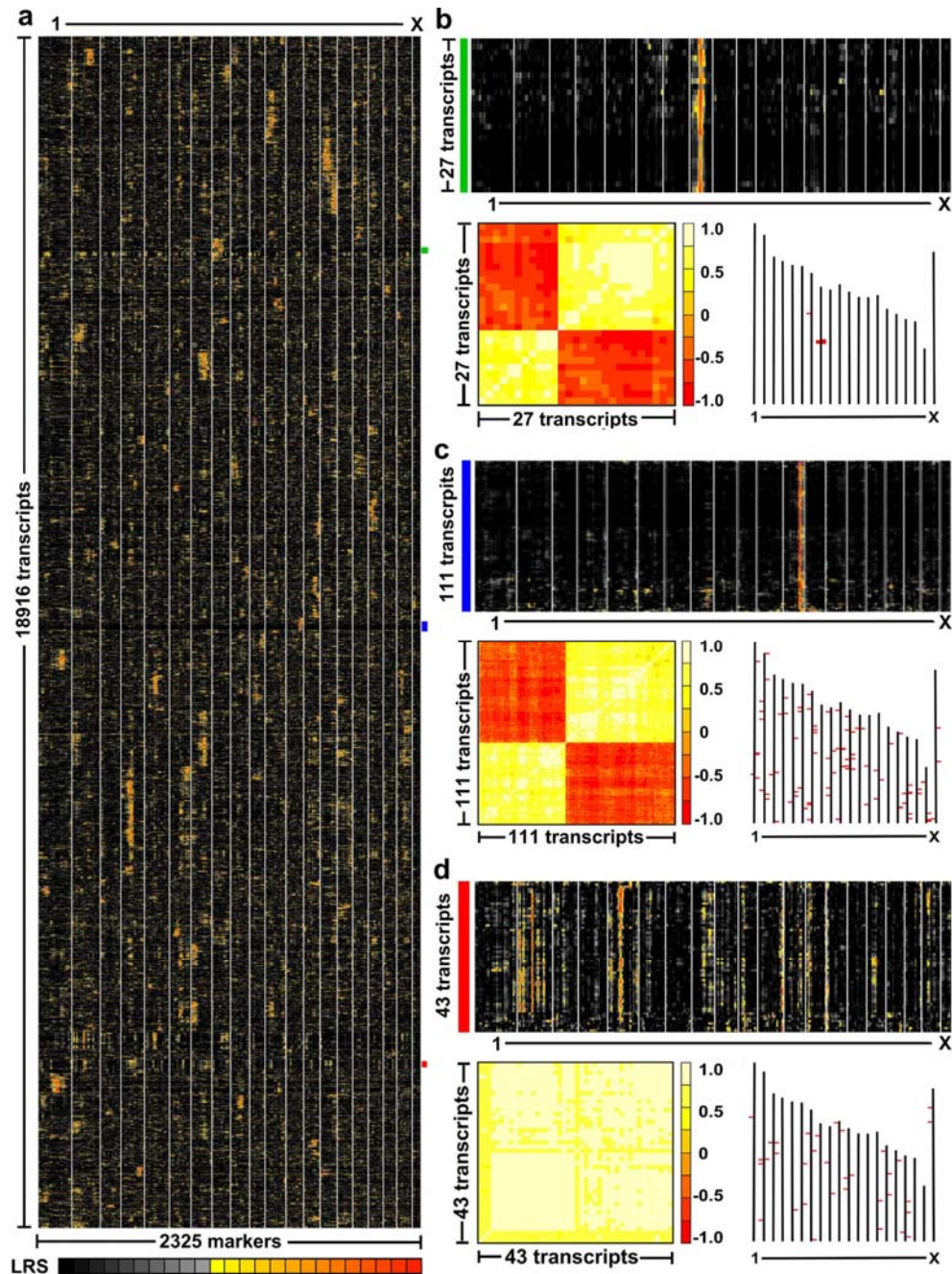


Figure 2. Genome-wide clustering of the genetic control of gene expression in liver. **a**, Hierarchical clustering diagram of all transcripts on the array by Likelihood Ratio Statistic (LRS) profile. Rows – transcripts on the microarray, columns – SNP markers used for the QTL analysis. Strength of the LRS values is depicted using a head map with black being the lowest LRS, and bright red – the highest. **b**, A zoomed-in view of the cluster of genes controlled by a single genetic locus on Chr8. An auto-correlation matrix of the measured expression values for the transcripts in the cluster and a plot of chromosomal location of these transcripts are shown below. **c**, A cluster of 111 transcripts controlled by one locus on Chr12. **d**, A cluster of 43 transcripts that are controlled by a complex set of loci on several chromosomes.

hence located under the maximal QTL. This region in the genome is a *master-regulator* since the majority of the transcripts are *trans*-regulated by this locus. Again, the correlation between expression levels of these transcripts is strongly dependent on the type of the parental allele that is present at this locus implying a common regulator of all 111 transcripts.

To further understand the genetic basis for the difference in the proportion of cis- and trans-regulated transcripts in simple clusters, 31 simple clusters comprised of at least 80% cis-regulated transcripts were examined. It was hypothesized that the clustering of these cis-regulated QTLs might be due to higher gene or SNP density as opposed to biological pathways. To determine why a large number of cis-regulated transcripts might be located in very small genetic loci gene density at each region was considered versus adjacent up- and down-stream regions. For gene density, it was found that the QTL cluster regions had 79.5 ± 70.0 (mean \pm SD) genes while the immediately adjacent up- and down-stream regions contained 45.0 ± 48.3 and 45.0 ± 44.2 genes, respectively. For SNP density, it was found that the QTL cluster regions had 4362 ± 3707 SNPs and the up- and down-stream regions had 2924 ± 2963 and 3318 ± 3957 SNPs, respectively. In this analysis, only those SNPs that differ between the C57BL/6J and DBA/2J strains in the Perlegen mouse SNP data set (mouse.perlegen.com/mouse/index.html) were considered. Thus, on the level of individual clusters, QTL clustering may be driven by a combination of gene and SNP density. Furthermore, the transcripts in these cis-regulated clusters do not appear to be enriched for any particular GO category. This indicates that the co-regulation of cis-controlled genes throughout the genome is not due to functional relatedness, but rather the high gene/SNP density in each region.

Lastly, a cluster of 43 transcripts (Figure 2d) that are controlled by a complex pattern of loci across multiple chromosomes (mean maximum LRS = 13.3) is shown. Not surprisingly, these genes are scattered around the genome. The pair-wise gene

expression correlation matrix for these transcripts shows that mRNA levels for these genes are highly positively correlated regardless of the allele type at each QTL.

The proportion of cis-eQTLs that had higher expression when the C57BL/6J allele is present was assessed. Following the criterion set out in Pierce et al.⁴³, a cis-eQTL was defined as a transcript that has a maximum QTL within +/- 5Mb of the transcript's location in the genome. At a genome wide p value of 0.05, 1,255 cis-eQTLs with 54.3% were found as having C57BL/6J high expression. At higher levels of statistical stringency 1,075 cis-eQTLs (p = 0.01) are found with 53.9% higher for the C57BL/6J allele and 867 cis-eQTLs (p = 0.001) with 53.3% higher for the C57BL/6J allele. The data presented by Doss et al.⁴² and the data from this study both show only a slight (54-56%) enrichment for C57BL/6J high cis-eQTLs when Agilent long oligo arrays, produced from the reference sequence of the C57BL/6J strain, are used. Consequently, much higher (72-75%) enrichment in the Pierce et al.⁴³ data set could be attributed to a low fidelity binding of short oligos used on Affymetrix arrays. These results show that long oligo array platforms are more suitable for eQTL analysis.

Mouse brain and liver transcriptomes show little overlap in genetic regulation of gene expression

Several recent reports identified a number of master-regulatory loci in other mouse tissues^{6,7,43}. Here, the mouse brain (forebrain) and liver transcriptome maps (Figure 3) are compared to uncover the similarities and differences in genetic regulation of gene expression across tissues. Both brain and liver contain genes that are strongly *cis*- or *trans*- regulated at single loci, or are regulated by multiple loci (Figures 3a and b, respectively). In the brain transcriptome, three distinct master-regulator trans-bands are located on chromosomes 1 and 2 (Figure 3c). In the liver transcriptome, the strongest trans-band is located on distal chromosome 12 (Figure 3d), a locus that does not appear

to be controlling expression of a large number of genes in the brain. While the liver and the brain both have a trans-band on chromosome 2 near 125 Mb, the two bands are not coincident. The liver trans-band lies at 119 MB and the brain at 135 Mb - a difference of 16 Mb.

Next, the QTL data for the liver and brain by selecting only those transcripts with a genome-wide p value ≤ 0.05 ⁴³ was filtered. This analysis identified 743 transcripts that have significant QTLs in both tissues (Figure 3e). It is interesting to note that 209 transcripts are regulated by the same genomic region in both the liver and the brain (genes that fall of a diagonal in Figure 3e). Gene Ontology-based⁴⁴ analysis of the biological processes that are significantly over-represented (Fisher's exact test $p < 0.05$ and 25% FDR) among these genes identified a number of significant categories that are co-regulated in both tissues (Figure 3f). Furthermore, among the 743 transcripts, the proportion of cis-regulated transcripts in the liver was 0.59 and in the brain it was 0.55. The difference is due to slight differences in the transcript location between the two data sets. Of the 209 transcripts that have the same QTL in the liver and the brain about 95% are cis-regulated. Collectively, this comparison indicates that important tissue-specific patterns of genetic control of gene expression can be elucidated by this approach and potentially form the basis for comparative analyses between tissues.

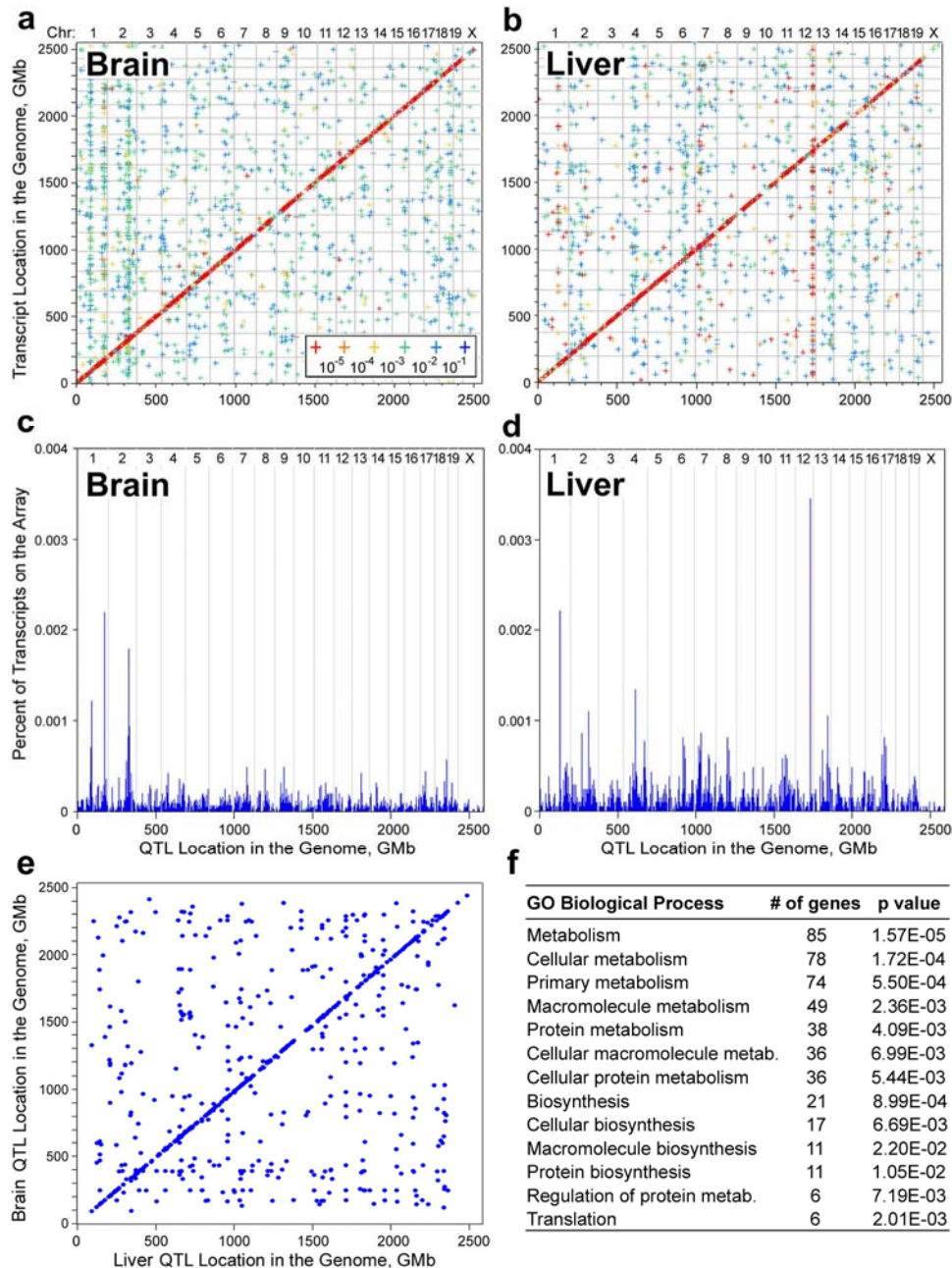


Figure 3. Tissue-specific transcriptome maps reveal differences and similarities in genetic regulation of gene expression. **a-b**, The brain⁴³ and liver (this study) transcriptome maps. The horizontal axis shows the genomic location of each genetic marker along the genome. The vertical axis shows the genomic location of each transcript probed by the microarray analysis. Each cross represents the location of the maximum QTL for a particular gene. Cis-regulated genes, where the QTL is co-located with the gene, fall along a 45 degree line. The vertical lines correspond to a locus with strong trans-control over many genes. The major loci of control differ markedly in the murine brain and liver. **c-d**, Histograms counting the proportion of transcripts on the array regulated at each marker. **e**, A comparative transcriptome map between the murine liver and brain. Significant QTLs in the liver and brain are plotted along the horizontal

and vertical axes, respectively. Points along the diagonal represent transcripts whose maximum QTL is the same in both tissues, indicating a similar mechanism of expression control. Points plotted off of the diagonal represent transcripts that are controlled by different loci in the two tissues. f, Gene Ontology analysis of the significantly enriched biological processes common between liver and brain in the mouse.

Chromosome 12 contains a strong liver-specific master-regulatory locus

The chromosome 12 locus regulates expression of 111 genes and is delineated by two SNP markers, rs13481620 at 98.47783 Mb and rs8273308 at 99.83812 Mb (Figure 4, average LRS = 42.6, genome-wide p value < 0.03 and 0.25 False Discovery Rate ⁴⁵). Surprisingly, the mean LRS for the transcripts that are cis-regulated (LRS = 31.4) at this locus is lower than that for the trans-regulated (LRS = 43.1) ones. It was hypothesized that this locus contains a gene that serves as a liver-specific “master regulator” of this chromosome 12 trans-band. To identify the candidate gene, genes in a 4 Mb region (98 - 102 Mb) around the two eQTL markers identified above were considered (Figure 4, lower panel).

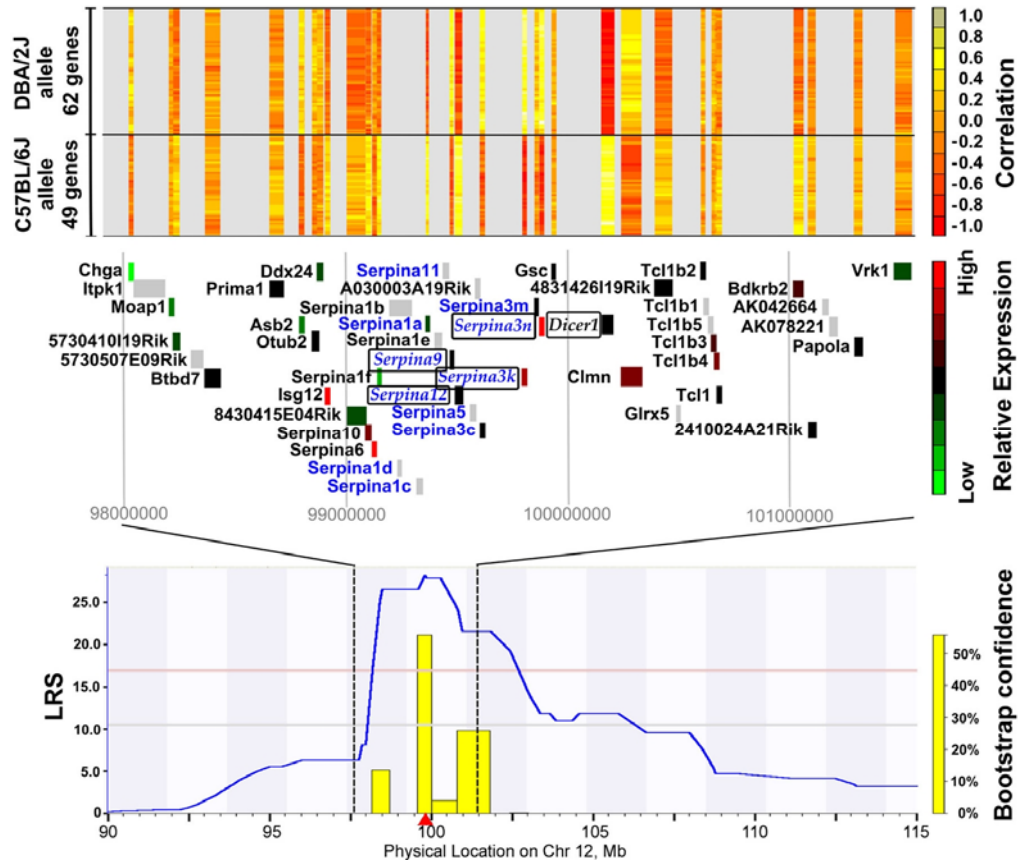


Figure 4. Chr12 locus is a master-regulator of gene expression in mouse liver. A zoomed-in view of the region on Chr12 that controls a large number of trans-regulated transcripts is shown in the bottom panel. The middle panel is a detail of the QTL peak from 98 Mb to 102 Mb showing the genes in this region. Genes labeled in blue text have non-synonymous coding SNPs. Genes in italics are cis-regulated. Genes in plain text are trans-regulated. The gene location marker is color coded based on relative expression in the mouse liver. Green indicates low expression; red indicates high expression; grey indicates that the gene is not represented on the microarray. The top panel shows the correlation between the expression of each gene located in this locus and the putative trans-regulated genes. Red indicates negative correlation; yellow indicates positive correlation; grey indicates no information. Genes in the upper part of the panel are highly expressed in strains with the DBA/2J allele at this Chr12 locus. Genes on the bottom of the panel are highly expressed in strains with the C57BL/6J allele is at this locus.

It was reasoned that the candidate gene that may be responsible for the variation in expression between the transcripts associated with this locus should satisfy the following properties: 1) be cis-regulated at the chromosome 12 locus or contain non-synonymous coding SNPs between parental alleles, 2) have median to high expression in

liver and 3) exhibit strong correlation in gene expression between the candidate gene and the trans-regulated transcripts when separated by parental allele at this locus. Five genes that are located in this region are cis-regulated: *Dicer1*, *Serpina3k*, *Serpina3n*, *Serpina9* and *Serpina12* (Figure 4, middle panel, genes identified in italics). Furthermore, a number of genes in this region, including *Serpina3k*, *Serpina3n*, *Serpina9*, and *Serpina12*, have known non-synonymous coding SNPs between the C57BL/6J and DBA/2J strains. *Serpina3k*, *Serpina3n*, *Dicer1*, *Serpina9* and *Serpina12*, among several other genes, also have median to high relative mRNA expression in liver. Lastly, when the strength of the correlation between expression of each transcript regulated by this locus and expression of all other transcripts located in the chromosome 12 locus is plotted (Figure 4, top panel), it is evident that *Dicer1*, *Serpina3k*, *Serpina3n*, *Serpina9*, *Serpina12* and four other transcripts have a clear separation (positive or negative correlation) according to the parental strain allele at this locus. Thus, it appears that any of the 5 cis-regulated genes at this locus: *Dicer1*, *Serpina3k*, *Serpina3n*, *Serpina9*, *Serpina12* is likely to be the candidate "master regulator" gene in liver.

Dicer1 is a logical candidate for the chromosome 12 locus since it is involved in post-transcriptional regulation of genes via cleavage of double-stranded RNAs (dsRNA) into ~20-25 base pair fragments. These fragments are then processed by the RNA-induced silencing complex, which recognizes sequences complementary to the dsRNA and prevents protein synthesis through either mRNA degradation or inhibition of translation⁴⁶. To date, no specific gene regulatory function has been proposed for *Dicer1* in the liver. To test the hypothesis that *Dicer1* is a master-regulator of gene expression in the liver, expression was compared between the chromosome 12 locus-regulated genes in livers of wild type and *Dicer1* heterozygous [*Dicer1* null mutation is embryonic lethal⁴¹] mice by quantitative real time PCR. As a negative control, a number of randomly selected genes that are not regulated by the chromosome12 locus were selected. Contrary to our

hypothesis, no consistent correlation was found between expression of *Dicer1* and chromosome 12 locus-regulated genes (data not shown) which suggests that *Dicer1* does not appear to be the master regulator at this locus. It should be noted, however, that *Dicer1* heterozygous mice may not be the most appropriate system for testing this hypothesis since *Dicer1* mRNA levels in heterozygotes are 134% of wild type levels and a small sample size ($n = 3$ per group) limits the power of the analysis ($p = 0.17$).

Next, other means of biological interpretation of the data were considered. Gene Ontology (GO) and transcription factor binding site analyses of the chromosome 12 locus trans-regulated genes were performed. GOMiner⁴⁷ examination of the 111 transcripts with maximum QTLs at the chromosome 12 locus identified significant enrichment for a single biological process category – cell surface receptor linked signal transduction ($p=8.74 \times 10^{-4}$). The genes from this category that are trans-regulated by the chromosome 12 locus are mainly olfactory receptors: *Bsf3*, *Rqcd1*, *Gpr50*, *Tcp10c*, *P2ry10*, *Olfr1403*, *Olfr1443*, *Olfr401*, *Olfr512*, *Olfr935*, *Olfr1341*, *Olfr341*, *Olfr656*, *Olfr1365*, *Mesp2*, *Met*, *Ltbp3*, *Fstl3*, *Centd2*, and *Rassf3*.

For the transcription factor binding site analysis, the trans-regulated transcripts at the chromosome 12 locus (LRS values greater than or equal to 30) were divided into two groups: those with high expression when either C57BL/6J or DBA/2J allele is present at the chromosome 12 locus. The National Cancer Institute's Advanced Biomedical Computing Center promoter analysis tool (grid.abcc.ncifcrf.gov/promoters.php) found no common transcription factor binding sites for the C57BL/6J list. LVC-Mo-MuLV and SV40.11 binding sites were identified as significant ($p = 9.766 \times 10^{-4}$) in the DBA/2J list. The oPossum³⁹ tool identified Freac-2 site as significant in both C57BL/6J and DBA/2J lists ($p = 6.026 \times 10^{-2}$ and $p = 1.823 \times 10^{-2}$, respectively), while ARNT ($p = 8.577 \times 10^{-2}$) and SOX-9 ($p = 9.159 \times 10^{-2}$) sites were also found to be common for DBA/2J allele-containing genes. The PAINT transcription factor tool⁴⁰ was also applied to the data and no

significant transcription factor binding sites between the two lists were found after FDR correction of the p-values. Similarly to our observation of the lack of a consistent signal for a transcription factor, Yvert et al.⁴⁸ and Kulp et al.⁴⁹ found that the genes in trans-regulated bands are not enriched for transcription factors or biological function. This suggests that the trans-regulated genes at the chromosome 12 locus have a complex mechanism of regulation that is yet to be discovered.

The other candidates for the master regulator at the chromosome 12 locus are Serpina3-family genes. These genes are the murine orthologs of human α 1-anti-chymotrypsin, a serine protease inhibitor. Serpina3n is an acute phase protein that increases 4 to 5-fold in inflammation and infection⁵⁰. In humans, α 1-anti-chymotrypsin is an inhibitor of neutrophil elastase, cathepsin G, mast cell chymase & pancreatic chymotrypsin⁵¹. In the mouse, *Seprin3a* targets are leukocyte elastase, cathepsin G and chymotrypsin⁵². While humans have one copy of α 1-anti-chymotrypsin at 14q32, the mouse has 14 copies at 12E1⁵³. α 1-Anti-chymotrypsin has been shown to be present in the amyloid plaques of Alzheimer's patients⁵³. Elzouki et al.⁵⁴ found an association between low plasma α 1-anti-chymotrypsin levels and propensity to contract the hepatitis B & C virus. A related protease inhibitor, serpin1 (α 1-antitrypsin) is involved in emphysema due to a failure to inhibit neutrophil elastase and cirrhosis due to an accumulation of serpin1 polymers in the hepatocytes^{51,55}. Although the overall structure is well conserved in the 14 member mouse *Serpina3* family, the reactive center loop is widely divergent, suggesting that these enzymes have function other than protease inhibition. Interestingly, human α 1-anti-chymotrypsin was reported to be able to bind to DNA and has been found to inhibit DNA polymerase and DNA primase *in vitro*⁵¹. Horvath et al.⁵² performed a detailed structural analysis of mouse SERPINA3N and found that it contains a DNA binding domain similar to one described in human α 1-anti-

chymotrypsin ⁵⁶. However, it was also reported that human α 1-anti-chymotrypsin binds to double stranded DNA without specificity to known DNA binding motifs ⁵⁶. It has also been shown that some serpins may require cofactors for activation ⁵¹ which raises the possibility that the behavior of the *Serpina3* family changes depending on the environment. Collectively, while there is no firm evidence for the role of *Serpina3* genes in regulation of gene expression, we posit that our data points to the potential novel role of this family of genes in regulating liver gene expression.

eQTL analysis facilitates the discovery of novel genotype-phenotype correlations

WebQTL contains comprehensive manually curated publicly available data for phenotypic and gene expression profiling of a number of recombinant inbred and F2 crosses in both mouse and rat along with the dense genetic marker maps for these strains. Thus, this data can be used to search for correlations between phenotypes, gene expression and genetic markers, i.e., to perform *in silico* genotype-phenotype association analysis. The inherent significance of the defined reference genetic populations, such as BXD RI strains, is in the ability to connect historical data generated in many laboratories to the exact genetic map of each strain. This provides an exceptional opportunity to add value and depth to the biological interpretation of the data from model organisms. Thus, even though the BXD RI panel of strains has not been used extensively to profile liver disease-specific responses, as compared to a wealth of behavioral phenotypes published over the years, it is not unreasonable to anticipate that more data will become available soon.

Here, to illustrate the power of combining genome-wide, liver expression profiling in a reference mouse panel with phenotype profiling, we identify several phenotypes that strongly correlate with the expression of liver transcription factor hepatocyte nuclear factor 4-gamma (*Hnf4g*, located on Chr3 at 3.620141 Mb) using standard tools available

in WebQTL. *Hnf4g* is trans-regulated by the chromosome 12 locus (Figure 5, left panel). Several of the BXD phenotypes are also regulated by this locus and there is strong correlation (Figure 5, right panel) between *Hnf4g* expression and the induction of serum IL-6 after TNF injection, lethality due to TNF injection and decreased body temperature after TNF injection⁵⁷. Furthermore, both the expression of *Hnf4g* and values of these phenotypes separate by parental allele at the location of the *Hnf4g* gene on proximal chromosome 3.

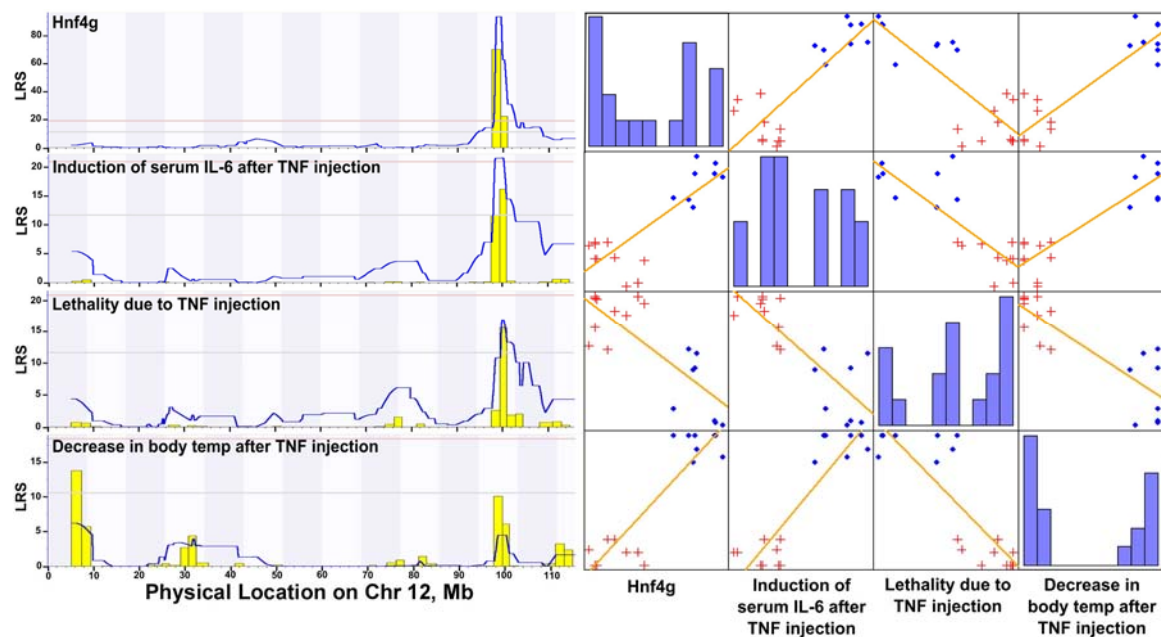


Figure 5. *In silico* discovery of gene expression to phenotype correlations using WebQTL. The left panel shows QTL interval maps on Chr12 for hepatocyte nuclear factor 4, gamma (*Hnf4g*) gene and three related phenotypes. The right panel shows correlation pair-wise correlation plots for *Hnf4g* expression and the phenotypes. Each dot represents the measurements for one strain. Blue dots indicate strains in which the C57BL/6J allele is present at the Chr12 (99.83Mb) locus. Red crosses represent strains with the DBA/2J allele at this locus. The blue histograms along the diagonal represent the distribution of all values of the phenotype.

Knowledge of the variability in basal gene expression in liver as a tool for selecting relevant strains to probe biology

Genetic engineering has been a powerful and useful tool in biomedical research. However, there are a number of instances when generating a knockout or knockin mouse is neither the best option, nor it is feasible. The process is both costly and lengthy. Furthermore, some genes are embryonic lethal while others are sufficiently redundant, thus limiting the ability to generate biologically meaningful data using genetic engineering approaches. Thus, we propose that understanding of the degree of variability in gene expression between strains in a reference population of mice may be used to model the potential biological effects of naturally occurring differences in mRNA levels between individuals. In fact, it is extremely rare that people are complete nulls for a particular gene, but the polymorphisms in certain genes are known to predispose the individuals to some exposures or lead to disease^{58,59}. Accordingly, the liver expression data in WebQTL may be utilized to select strains that differ in basal mRNA level of genes of interest and then used for phenotypic studies that are designed to test the role of the genes in a particular phenotype.

The flavin-containing mono-oxygenases are a class of phase I enzymes that oxidize organic nitrogen and sulfur containing compounds such as cimitidine, methimazole and nicotine⁶⁰⁻⁶². Fmo3 is the most common isoform in the human liver⁶³. While no sex differences in human activity of Fmo3 have been shown, in mice *Fmo3* expression has been shown to be much higher in females^{63,64}. Basal expression of *Fmo3* varies widely across the BXD strains (Figure 6, top panel). We confirmed this by running quantitative real time PCR on strains with high, medium and low expression of *Fmo3* in 9 out of 40 strains (Figure 6, lower panel). The correlation between the expression measured by microarray and PCR is strong ($R^2 = 0.74$). Thus, WebQTL can be used to query the expression database to find genes with the highest differential

expression across strains to generate a potential list of candidate strains for hypothesis testing. Since *Fmo3* expression varies widely across BXD strains and mutations in this gene have been implicated in trimethylaminuria⁵⁸, a disease in which trimethylamine is not metabolized but is excreted in the breath and sweat, leading to a persistent fishy body odor, we suggest that BXD strains with high or low relative expression of *Fmo3* may be used to model this disease. Similar logic may be used to test other genes without the cost and time of generating knockout animals.

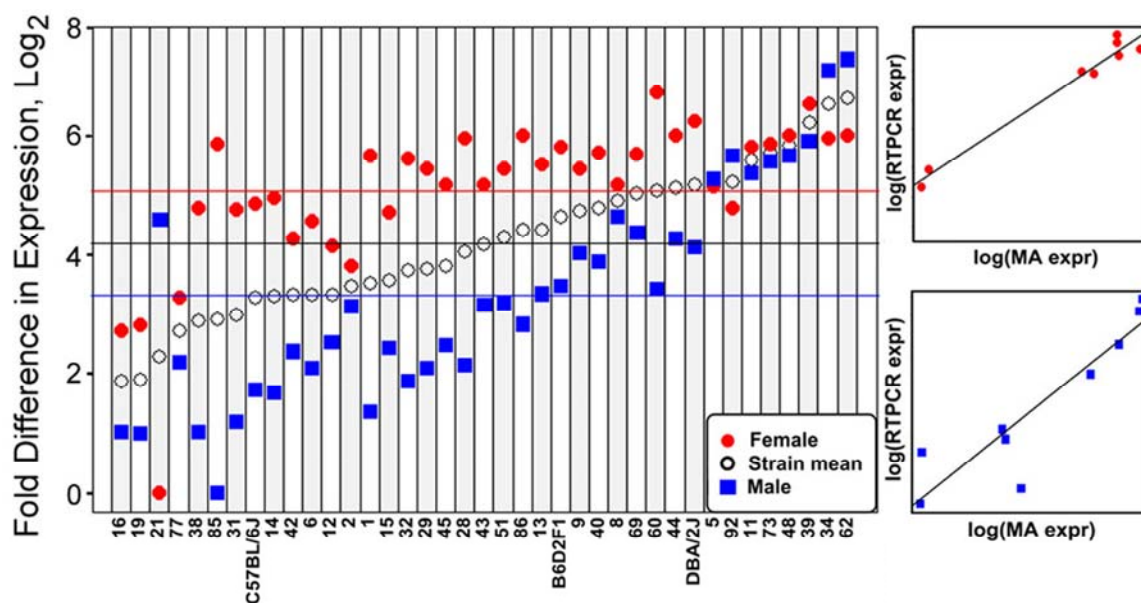


Figure 6. WebQTL-assisted strain selection for phenotypic profiling. The main panel shows relative expression of *Fmo3* (flavin monooxygenase 3) across BXD strains measured by microarray. Red circles are values in females; blue squares – in males and the strain means are shown as black circles. The lower panels show relative *Fmo3* expression in select strains as measured by microarray vs. quantitative real time PCR. The expression of *Fmo3* in BXD21 strain was set to 0.

CHAPTER 5

Conclusions and Future Directions

In conclusion, this study describes a new public resource that will facilitate our understanding of the genetic regulation of gene expression in liver. We describe several genetic loci that control expression of large numbers of genes. By using eQTL mapping, we identified the *Serpina3* family of genes as potential novel master regulators of transcription in the liver. By comparing the liver and brain transcriptome maps, we highlighted tissue-specific differences in regulation of gene expression. Finally, this study demonstrates how this publicly available data may be used to infer genotype-phenotype correlations, generate testable hypothesis, and select mouse strains for further testing based on the genetically-determined differences of expression of the key genes.

What gene(s) regulates the Chr 12 trans-QTL band?

The work presented above regarding the chromosome 12 trans-regulated band could be further extended by examining the corresponding set of genes in a separate panel of inbred mice. Data has recently been generated in the Rusyn laboratory in which the constitutive gene expression of a panel of 36 strains of inbred strains was measured. Each strain consists of 2 or 3 microarrays from each strain. Although all samples in this inbred data set are male, the chromosome 12 eQTLs have lower p-values in the BXD male samples and this single sex analysis will not present a problem. These 36 inbred strains are part of the Novartis genotyping project. This project has genotyped 71 inbred

strains (including the 36 strains above) at over 150,000 markers. The increased density as well as the large number of strains should provide both confirmation of the genes truly regulated by the chromosome 12 locus and a narrower QTL to permit inference of the regulatory gene(s).

Once we have selected a small set of putative regulatory genes, cultures of primary hepatocytes from a subset of the BXD strains will be prepared. First, the stability of the eQTL trans-band *in vitro* will be determined. The expression of genes with high expression when the C57BL/6J allele is present at the chromosome 12 locus should remain higher than the expression of genes that are low when the DBA/2J locus is present. RTPCR will be used to measure a subset of the trans-regulated genes that appear in both the BXD and inbred data sets. Once this is established, the levels of mRNA of the putative regulatory genes will be knocked down using RNAi technology^{65,66}. RTPCR will then be used to measure the expression of the expected regulator and a subset of positively and negatively correlated genes that are regulated by this locus. The hypothesis is that when the regulatory gene(s) is knocked down, the expression levels of the corresponding regulated genes will increase or decrease depending on the direction of their correlation with the regulatory gene.

What role do microRNAs play in the regulation of gene expression?

The use of microarrays to measure gene expression leads naturally to a focus on polymorphism in genes and their promoter regions as the causative explanation for the variation in gene expression. However, recent evidence shows that RNA interference influences the levels of mRNA transcripts in the cell⁴⁶. Briefly, double stranded RNA (dsRNA) is transcribed in the nucleus, exported to the cytoplasm and cleaved into 21 to 23 nucleotide long small interfering RNAs (siRNA) by Dicer. One strand of the dsRNA

associates with the RNA-induced silencing complex (RISC) which recognizes the complementary RNA sequences and targets mRNA for destruction⁴⁶. There are other proposed mechanisms by which RNA interference (RNAi) affects expression levels. The siRNA may bind to the mRNA and inhibit translation without actual destruction of the mRNA. siRNA may also prevent transcription by binding to DNA at promoter sites or affecting methylation status⁴⁶.

Microarrays have been developed^{67,68} that measure the levels of microRNA expression. There are about 385 known mouse microRNAs currently in the Sanger microRNA data base^{69,70}. Working with collaborators at Oak Ridge National Labs, the expression of known microRNAs will be measured in the livers of the BXD mouse panel and these will be correlated with gene expression and used to find the causative polymorphisms underlying individual QTLs⁷¹. There is one important caveat. If siRNA prevents transcription or acts through transcriptional degradation and there is a polymorphism that affects siRNA binding, then it should be possible to measure differences in target transcript abundances. However, if the siRNA operates through translational repression, then transcript levels measured in the BXD panel will not be altered and no QTL will be evident⁷¹.

What is the role of sex on constitutive gene expression?

The effect of sex on the control of gene expression was not evaluated in this study. Although there is only one microarray per sex per strain, this type of analysis could provide preliminary insight and testable hypotheses. The location of the trans-regulated bands differs by sex. A transcriptome map of the male data (not shown) reveals trans-bands on proximal chromosome 7 and distal chromosome 12. The same plot with the female data chromosomes 1, 4, 6, 11, 12, 15 & 19. Clusters of co-regulated genes that

differ in expression by sex were compiled by ORNL using a new clique finding algorithm⁷². These clusters will be examined for biological relevance using the Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, and Ingenuity Pathway Analysis.

In summary, genetical genomics is a powerful approach to understanding the genetic basis of gene expression. When gene expression is a heritable trait, panels of recombinant inbred mice can be used to dissect the networks of co-regulated genes that vary due to genetic polymorphisms across strains. These networks and the loci that regulate their expression offer insight into the working of the liver and provide data that can be correlated with phenotypic measurements for mechanistic hypothesis generation.

Table 1 : Microarray Experiment Batch Design

		COMPLEMENT 1							
		Batch 1		Batch 2		Batch 3		Batch 4	
Strain		M	F	M	F	M	F	M	F
C57BL/6J		614	5	8	5	506	5		5
		614		8		506			5
							605		823
DBA/2J			1	4	1	509	1		1
					3				607
B6D2F1		609		603	601	702	602		
BXD1		9							
BXD12			13						
BXD11				12					
BXD13					16				
BXD14						17			
BXD15							19		
BXD16								803	
BXD23		24							704
BXD34			43						
BXD21				21					
BXD24									25
BXD33									
BXD9							70		
BXD19									804
BXD28		29							
BXD29			30						
BXD31				34					
BXD32					36			701	
BXD36						46			
BXD38							52		
BXD39		54							
BXD40			59					60	
BXD42				63					
BXD5					65				
BXD6						69			
BXDA23	BXD86								819
BXD2									
BXD8			613					817	
BXD69									813
BXD73								816	
BXD92									821
BXDA10	BXD77	500+501							
BXDAP11	BXD51							77	
BXDAP15	BXD62			79					812
BXDAP19	BXD43				81				
BXDAP5	BXD60					86			
BXDAP6	BXD44						87		
BXDAP8	BXD48								
BXDA22	BXD85								818
BXDAP12	BXD45								

		Batch 5		Batch 6		Batch 7		Batch 8	
Strain		M	F	M	F	M	F	M	F
C57BL/6J			5	507	5		5		5
			5		5		5		5
DBA/2J			1	510	1	511	1		1
B6D2F1									801
BXD1			11						
BXD12		14							
BXD11									703
BXD13		15							
BXD14									610
BXD15		18							
BXD16									802
BXD23									
BXD34				42					
BXD21					20				
BXD24				26					
BXD33					40				809
BXD9				71					
BXD19								805	
BXD28		28							
BXD29						31			
BXD31					32				
BXD32									
BXD36									48
BXD38						51			
BXD39									56
BXD40									
BXD42							62		
BXD5						66			
BXD6							68		
BXDA23	BXD86							901	
BXD2					612			611	
BXD8									
BXD69								814	
BXD73									815
BXD92								822	
BXDA10	BXD77		499						
BXDAP11	BXD51								902
BXDAP15	BXD62								
BXDAP19	BXD43					82			
BXDAP5	BXD60		84						
BXDAP6	BXD44			502+503					
BXDAP8	BXD48			512					92
BXDA22	BXD85							903	
BXDAP12	BXD45		515					811	

Table 2 : BXD strain information

Strain Name	Sex
C57BL/6J	M and F
DBA/2J	M and F
B6D2F1	M and F
BXD1	M and F
BXD2	M and F
BXD5	M and F
BXD6	M and F
BXD8	M and F
BXD9	M and F
BXD11	M and F
BXD12	M and F
BXD13	M and F
BXD14	M and F
BXD15	M and F
BXD21	M and F
BXD23	F
BXD24	M
BXD28	M and F
BXD29	M and F
BXD31	M and F
BXD32	M and F
BXD33	F
BXD34	M and F
BXD38	M and F
BXD39	M and F
BXD40	M and F
BXD42	M and F

BXD43	M and F
BXD44	M and F
BXD45	M and F
BXD48	M and F
BXD51	M and F
BXD60	M and F
BXD62	M and F
BXD69	M and F
BXD73	M and F
BXD77	M and F
BXD85	M and F
BXD86	M and F
BXD92	M and F

References

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al. The sequence of the human genome. *Science* 2001;291:1304-1351.
2. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;420:520-562.
3. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308-311.
4. Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Lavolette JP et al. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet* 2000;24:381-386.
5. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 2003;422:297-302.
6. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* 2005;37:233-242.
7. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* 2005;37:225-232.
8. Schadt EE. Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. *Curr Opin Biotechnol* 2005;16:647-654.
9. Holt MP, Ju C. Mechanisms of drug-induced liver injury. *AAPS J* 2006;8:E48-E54.
10. Watkins PB, Seeff LB. Drug-induced liver injury: summary of a single topic clinical research conference. *Hepatology* 2006;43:618-631.
11. Crebelli R, Carere A. Genetic toxicology of 1,1,2-trichloroethylene. *Mutat Res* 1989;221:11-37.
12. Klaassen CD, Liu J. Metallothionein transgenic and knock-out mouse models in the study of cadmium toxicity. *J Toxicol Sci* 1998;23 Suppl 2:97-102.
13. Melnick RL, Huff J. 1,3-butadiene: Toxicity and carcinogenicity in laboratory animals and in humans. *Rev Environ Contamin Toxicol* 1992;124:111-144.
14. Harbrecht BG, Billiar TR. The role of nitric oxide in Kupffer cell-hepatocyte interactions. *Shock* 1995;3:79-87.
15. Clayton TA, Lindon JC, Cloarec O, Antti H, Charuel C, Hanton G et al. Pharmacometabonomic phenotyping and personalized drug treatment. *Nature* 2006;440:1073-1077.

16. Wade CM, Daly MJ. Genetic variation in laboratory mice. *Nat Genet* 2005;37:1175-1180.
17. Paigen K, Eppig JT. A mouse phenome project. *Mamm Genome* 2000;11:715-717.
18. Svenson KL, Von Smith R, Magnani PA, Suetin HR, Paigen B, Naggert JK et al. Multiple Trait Measurements in 43 Inbred Mouse Strains Captures the Phenotypic Diversity Characteristic of Human Populations. *J Appl Physiol* 2007.
19. Wang X, Paigen B. Genetics of variation in HDL cholesterol in humans and mice. *Circ Res* 2005;96:27-42.
20. Grisel JE, Metten P, Wenger CD, Merrill CM, Crabbe JC. Mapping of quantitative trait loci underlying ethanol metabolism in BXD recombinant inbred mouse strains. *Alcohol Clin Exp Res* 2002;26:610-616.
21. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 1989;121:185-199.
22. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science* 2002;296:752-755.
23. Waring JF, Jolly RA, Ciurlionis R, Lum PY, Praestgaard JT, Morfitt DC et al. Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicol Appl Pharmacol* 2001;175:28-42.
24. Powell CL, Kosyk O, Ross PK, Schoonhoven R, Boysen G, Swenberg JA et al. Phenotypic anchoring of acetaminophen-induced oxidative stress with gene expression profiles in rat liver. *Toxicol Sci* 2006;93:213-222.
25. Taylor BA. Recombinant inbred strains. In: Lyon ML, Searle AG, editors. *Genetic Variants and Strains of the Laboratory Mouse*. Oxford, UK: Oxford University Press, 1989: 773-796.
26. Farrall M. Quantitative genetic variation: a post-modern view. *Hum Mol Genet* 2004;13 Spec No 1:R1-R7.
27. Morse HCl, editor. *Origins of inbred mice: Proceedings of a workshop*. 78 Feb 14; Academic Press, 1978.
28. Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet* 2004;5:7.
29. Treadwell JA. Integrative strategies to identify candidate genes in rodent models of human alcoholism. *Genome* 2006;49:1-7.
30. Gill K, Liu Y, Deitrich RA. Voluntary alcohol consumption in BXD recombinant inbred mice: relationship to alcohol metabolism. *Alcohol Clin Exp Res* 1996;20:185-190.

31. Bigelow SW, Nebert DW. The murine aromatic hydrocarbon responsiveness locus: a comparison of receptor levels and several inducible enzyme activities among recombinant inbred lines. *J Biochem Toxicol* 1986;1:1-14.
32. Lee GH, Bennett LM, Carabeo RA, Drinkwater NR. Identification of hepatocarcinogen-resistance genes in DBA/2 mice. *Genetics* 1995;139:387-395.
33. Davis RC, Schadt EE, Cervino AC, Peterfy M, Lusis AJ. Ultrafine mapping of SNPs from mouse strains C57BL/6J, DBA/2J, and C57BLKS/J for loci contributing to diabetes and atherosclerosis susceptibility. *Diabetes* 2005;54:1191-1199.
34. Nishina PM, Wang J, Toyofuku W, Kuypers FA, Ishida BY, Paigen B. Atherosclerosis and plasma and liver lipids in nine inbred strains of mice. *Lipids* 1993;28:599-605.
35. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU et al. Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2005;2:351-356.
36. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002;30:e15.
37. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. *Genetics* 1994;138:963-971.
38. Storey JD. A direct approach to false discovery rates. *J R Statist Soc B* 2002;64:479-498.
39. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP et al. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res* 2005;33:3154-3164.
40. Vadigepalli R, Chakravarthula P, Zak DE, Schwaber JS, Gonye GE. PAINT: a promoter analysis and interaction network generation tool for gene regulatory network identification. *OMICS* 2003;7:235-252.
41. Bernstein E, Kim SY, Carmell MA, Murchison EP, Alcorn H, Li MZ et al. Dicer is essential for mouse development. *Nat Genet* 2003;35:215-217.
42. Doss S, Schadt EE, Drake TA, Lusis AJ. Cis-acting expression quantitative trait loci in mice. *Genome Res* 2005;15:681-691.
43. Peirce JL, Li H, Wang J, Manly KF, Hitzemann RJ, Belknap JK et al. How replicable are mRNA expression QTL? *Mamm Genome* 2006;17:643-656.
44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25-29.
45. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 1995;B 57:289-300.

46. Hannon GJ. RNA interference. *Nature* 2002;418:244-251.
47. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 2003;4:R28.
48. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN et al. Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat Genet* 2003;35:57-64.
49. Kulp DC, Jagalur M. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics* 2006;7:125.
50. Kalsheker N, Morley S, Morgan K. Gene regulation of the serine proteinase inhibitors alpha1-antitrypsin and alpha1-antichymotrypsin. *Biochem Soc Trans* 2002;30:93-98.
51. Janciauskiene S. Conformational properties of serine proteinase inhibitors (serpins) confer multiple pathophysiological roles. *Biochim Biophys Acta* 2001;1535:221-235.
52. Horvath AJ, Irving JA, Rossjohn J, Law RH, Bottomley SP, Quinsey NS et al. The murine orthologue of human antichymotrypsin: a structural paradigm for clade A3 serpins. *J Biol Chem* 2005;280:43168-43178.
53. Forsyth S, Horvath A, Coughlin P. A review and comparison of the murine alpha1-antitrypsin and alpha1-antichymotrypsin multigene clusters with the human clade A serpins. *Genomics* 2003;81:336-345.
54. Elzouki AN, Verbaan H, Lindgren S, Widell A, Carlson J, Eriksson S. Serine protease inhibitors in patients with chronic viral hepatitis. *J Hepatol* 1997;27:42-48.
55. Perlmutter DH. Pathogenesis of chronic liver injury and hepatocellular carcinoma in alpha-1-antitrypsin deficiency. *Pediatr Res* 2006;60:233-238.
56. Naidoo N, Cooperman BS, Wang ZM, Liu XZ, Rubin H. Identification of lysines within alpha 1-antichymotrypsin important for DNA binding. An unusual combination of DNA-binding elements. *J Biol Chem* 1995;270:14548-14555.
57. Libert C, Wielockx B, Hammond GL, Brouckaert P, Fiers W, Elliott RW. Identification of a locus on distal mouse chromosome 12 that controls resistance to tumor necrosis factor-induced lethal shock. *Genomics* 1999;55:284-289.
58. Dolphin CT, Janmohamed A, Smith RL, Shephard EA, Phillips IR. Missense mutation in flavin-containing mono-oxygenase 3 gene, FMO3, underlies fish-odour syndrome. *Nat Genet* 1997;17:491-494.
59. Rudnick DA, Perlmutter DH. Alpha-1-antitrypsin deficiency: a new paradigm for hepatocellular carcinoma in genetic liver disease. *Hepatology* 2005;42:514-521.
60. Ziegler DM. Recent studies on the structure and function of multisubstrate flavin-containing monooxygenases. *Annu Rev Pharmacol Toxicol* 1993;33:179-199.

61. Cashman JR, Park SB, Berkman CE, Cashman LE. Role of hepatic flavin-containing monooxygenase 3 in drug and chemical metabolism in adult humans. *Chem Biol Interact* 1995;96:33-46.
62. Ripp SL, Overby LH, Philpot RM, Elfarra AA. Oxidation of cysteine S-conjugates by rabbit liver microsomes and cDNA-expressed flavin-containing mono-oxygenases: studies with S-(1,2-dichlorovinyl)-L-cysteine, S-(1,2,2-trichlorovinyl)-L-cysteine, S-allyl-L-cysteine, and S-benzyl-L-cysteine. *Mol Pharmacol* 1997;51:507-515.
63. Ripp SL, Itagaki K, Philpot RM, Elfarra AA. Species and sex differences in expression of flavin-containing monooxygenase form 3 in liver and kidney microsomes. *Drug Metab Dispos* 1999;27:46-52.
64. Falls JG, Blake BL, Cao Y, Levi PE, Hodgson E. Gender differences in hepatic expression of flavin-containing monooxygenase isoforms (FMO1, FMO3, and FMO5) in mice. *J Biochem Toxicol* 1995;10:171-177.
65. Lin SL, Ying SY. Gene silencing in vitro and in vivo using intronic microRNAs. *Methods Mol Biol* 2006;342:295-312.
66. Kuhn R, Streif S, Wurst W. RNA interference in mice. *Handb Exp Pharmacol* 2007;149-176.
67. Thomson JM, Parker J, Perou CM, Hammond SM. A custom microarray platform for analysis of microRNA gene expression. *Nat Methods* 2004;1:47-53.
68. Hammond SM. RNAi, microRNAs, and human disease. *Cancer Chemother Pharmacol* 2006;58 Suppl 1:s63-s68.
69. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006;34:D140-D144.
70. Griffiths-Jones S. The microRNA Registry. *Nucleic Acids Res* 2004;32:D109-D111.
71. Bao L, Zhou M, Wu L, Lu L, Goldowitz D, Williams RW et al. PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits. *Nucleic Acids Res* 2007;35:D51-D54.
72. Voy BH, Scharff JA, Perkins AD, Saxton AM, Borate B, Chesler EJ et al. Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS Comput Biol* 2006;2:e89.